# DATA-DRIVEN PHRASING FOR SPEECH SYNTHESIS IN LOW-RESOURCE LANGUAGES

*Alok Parlikar and Alan W Black*

Language Technologies Institute, Carnegie Mellon University, USA
aup, awb @cs.cmu.edu

## ABSTRACT

We present an approach to build phrase break prediction models when synthesizing text in low resource languages. This method allows building models without depending on the availability of part of speech taggers, or corpus with hand annotated breaks. We use the same speech data used for building a synthetic voice, to deduce acoustic phrase breaks. We perform unsupervised part of speech induction over a small text corpus in the language at hand. We use these tags and train a grammar based phrasing model. In this paper, we show results for the languages: English, Portuguese and Marathi, which suggest that we can quickly build very reasonable phrasing models for new languages using very little data.

***Index Terms***— Speech Synthesis, Phrase Break Prediction, Low Resource Languages

## 1. INTRODUCTION

Predicting prosodic phrases (phrase breaks) is an essential step during speech synthesis, because other prosodic models depend on it. The problem of phrasing can be thought of as predicting whether a break should be synthesized at each word boundary, or not. Phrase break prediction models are typically trained on standard corpora that contain hand annotated breaks. For example, the Festival[1] system uses a Part-of-Speech (POS) based model[2] trained on the MARSEC[3] data for English voices.

POS models trained on hand annotated phrase breaks can yield very reasonable phrasing models. However, many of the languages we build synthetic voices for do not have rich linguistic resources. Annotating text with phrase breaks is a laborious process, and annotating text to train POS taggers is even more expensive.

The default phrasing model for new languages simply uses punctuation to insert breaks during synthesis. Indeed, commas, semi colons, hyphens and full stops are very good indicators of breaks. However, we often get to synthesize text which does not contain punctuation. Some languages do not use punctuation in text in the manner that English does. And in some situations, we have to synthesize spoken language translations, where the output of Automatic Speech Recognition and Machine Translation may lack punctuation. The default punctuation-based phrasing model then results in long synthesized utterances without breaks, which sounds very unnatural—breathless, and with inappropriate pitch contours, for example.

In this paper, we suggest a data driven approach that can build good phrasing models without dependence on the availability of linguistic resources. Instead of requiring hand annotated phrase breaks, we derive phrase breaks from speech data—the same data that we use to build our synthetic voice. In place of using POS tags, we use a small text corpus and perform unsupervised POS induction. We build a CART model over the acoustically derived phrases, using these Induced-Part-of-Speech (IPOS) tags as features.

We present our work on three languages here: English, Portuguese, and Marathi. In Section 2, we present the details of resources available for each language. In Section 3, we describe our methodology to build the phrasing models. Before we describe our results, we introduce in Section 4 the design of our objective and subjective evaluation. In Section 5, we present our subjective and objective results on the three languages.

## 2. LANGUAGES AND AVAILABLE RESOURCES

We wanted to carry out experiments on languages that differ in families as well as amount of linguistic resources available. We chose English, European Portuguese and Marathi for this work. English is a Germanic language with rich set of linguistic tools. Portuguese is a Romance language and has many linguistic resources available in general. Marathi is an Indo-Aryan language spoken in India, and all we had access to was a text corpus.

Our English voice was trained on the F2B corpus (about 55 mins of speech) from the Boston University Radio News Corpus (BURNC)[4]. We had an English POS tagger available within Festival. We induced IPOS tags over 50000 sentences taken from the English side of the Europarl[5] corpus. For running listening tests, we randomly selected 25 long utterances from the F1A corpus in BURNC.

We built our Portuguese voice from about an hour of speech of recordings of a male news broadcaster from Portuguese national TV. We did not have access to a Portuguese POS tagger. We did have a lexicon that provides part of speech for known words, but does not disambiguate multiple possible POS based on context. We used 50000 sentences from the English-Portuguese Europarl corpus and induced IPOS tags for Portuguese. For running listening tests, we selected 15 long utterances from online Portuguese newspapers.

We only had a text corpus available for Marathi. This was a collection of news published in the E-Sakal newspaper. The corpus was collected at the Center for Indian Language Technology at IIT Bombay. We recorded about half an hour of speech to both build a synthetic voice, and build phrasing models. There was no POS tagger or lexicon available. We used 50000 sentences from the text corpus to induce IPOS tags for Marathi. For listening tests, we selected 15 long utterances from this same corpus.

Phrase prediction is an easier problem when text is well punctuated. In order to simulate the harder (and more important) case when punctuation is not available to us during synthesis, we stripped all our corpora for punctuation within utterances for all languages. We let the sentence final punctuation remain in text.

Note that for all three languages, we ran IPOS learning only on 50000 sentences. We did have access to much larger text corpus in all languages, but we decided to using a corpus of this size to make

sure our technique works when for a new language we might not have hundreds of thousands of lines of text available.

## 3. BUILDING PHRASING MODELS

We used our Grammar Based Approach[6] to build phrasing models in this work. This algorithm requires two things: (i) A corpus with labels of which word boundaries are breaks, and (ii) POS tags for words, or something similar. We shall see in Sections 3.1 and 3.2 how we can satisfy these requirements. In Section 3.3, we shall then quickly summarize our modeling method.

### 3.1. Acoustically Derived Phrase Breaks

Because we do not have a corpus with hand annotated breaks, we derive the breaks from our speech data. We force align the speech with its transcription using an HMM tool[7]. This tool allows us to bootstrap alignments just from speech and its transcripts for any language. Word boundaries which get aligned to short silences get marked as having a break. Other word boundaries are marked to have no break. Since our speech corpus typically consists of a few hundred utterances, we get annotations for breaks over a few hundred sentences to learn from.

### 3.2. Unsupervised POS Induction

If we do not have a POS tagger available, we can train an unsupervised model to induce POS tags. We used the Ney-Essen clustering algorithm[8] implemented in the POS-Induction tool[9] and ran it over the text corpus for each language. This algorithm iteratively improves the likelihood of a given clustering by moving each word from its current cluster to a cluster that will maximize the increase in likelihood.

We only clustered words that appeared in our corpus over 1000 times, and grouped them into 16 clusters. We used these clusters without any modifications and plugged them into Festival so that an IPOS tag is available for words at synthesis time. If a tag is not available for a certain word at run time, we assigned it a default tag called "content".

### 3.3. Grammar Based Phrasing Model

Once we have the acoustic phrases and the IPOS tags ready for a data set, we build a grammar based phrasing model[6]. We use the phrase break information to induce bracketing over text. For example: *(There are) (five hundred students) (in the room)*. These brackets represent a concept similar to constituency in traditional linguistics. However, in this case, we only care about prosodic phrases and prosodic constituents may not be valid linguistic constituents. Once we have this bracketing structure for our training data, we train a Stochastic Context Free Grammar (SCFG) that can be used to induce similar bracketing over new text.

Given a new utterance, we parse it with our SCFG. We then use a CART model that combines word level features with syntactic features over the parse, and predict whether each word boundary is a break or a non-break.

## 4. EVALUATION METHODS

In this section, we first describe the different objective metrics we use to compare the different models we built. We then describe our setup for subjective listening tests and how we interpret the subjective results.

### 4.1. Objective Evaluation Metrics

For data held out from the training corpus, we have information, from the acoustically derived phrases, about where the breaks should be. We can run our models over those utterances and find out how close they are in break prediction. Predicting a break at the end of an utterance is a trivial task, and hence we exclude all sentential breaks when counting how good our models are. Here are the objective metrics we use for evaluating our models:

#### *4.1.1. Measuring Accuracy (F-1)*

We can measure how accurate our break prediction was. We should be predicting as many breaks as there are in our reference (high recall), and yet not predicting breaks in wrong places (high precision). We can thus calculate the F-1 measure[10] and evaluate our models. A better model would get higher F-1 score, and the truly accurate model would have a score of $1.00$.

#### *4.1.2. Comparing Phrase Length Histograms*

Given a text utterance, different people might phrase it in different ways. Instead of comparing whether we got each of the breaks correct, we could look at the length profile of phrases predicted. Phrase length is the number of words between two consecutive breaks. We can build a histogram of the phrase lengths predicted by a model. We can then compare that to the histogram of phrase lengths found in the acoustically derived phrases. By measuring the distance between the histograms of our model, and the reference, we can evaluate how close our model was to the truth. We use the two commonly used metrics to compare histograms: the L2 distance and the Earth Mover's Distance (EMD). A fully accurate model would get both the L2 and EMD distances to be $0.00$, and a model with lower score would be deemed better.

- L2 Distance
  If we represent two histograms as single dimensional vectors, then the L2 distance is simply the Euclidean (or L2) distance between the two vectors. If $\mathbf{a}$ and $\mathbf{b}$ are two histogram vectors, then
  $$D_{L2}(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_i (a_i - b_i)^2}$$
- Earth Mover's Distance
  The Earth Mover's Distance[11] (EMD) between two distributions is proportional to the minimum amount of *work* required to change one distribution into the other. We normalize our histograms so that they represent probability distributions of the phrase lengths. For distributions in one dimension, such as our histograms, it has been proved[12] that the EMD between two distributions is the area between the graphs of the cumulative distributions.

### 4.2. Subjective Evaluation

We performed subjective listening tests for English, Portuguese and Marathi to compare phrasing models. Listening tests were set up as a web based A-B task. Two phrasing models were compared at a time. An utterance was synthesized by both models and both versions presented to the participants. After listening to the two versions, participants were asked to mark which version they preferred. They could

also choose a neutral option if they couldn't pick one over the other. Utterances were presented in random order, and the two versions of each utterance were also randomly ordered on the web page.

For English, we used Amazon Mechanical Turk (MTurk) to run the listening task. We split 25 utterances into sets of 5x5. Each set was presented as an individual HIT. We allowed 10 workers per HIT. Thus, we had 50 tasks, and 5 utterances each, giving us 250 data points for comparison. We discarded responses by few workers on MTurk since they had finished the task too quickly, and their responses would have been spam.

For Portuguese and Marathi we could not reliably use MTurk for the listening task. We requested volunteer native speakers of the languages to perform the task. Majority of our Portuguese participants did the task over the web from Portugal, and similarly majority Marathi tests were taken in India. We had about 100 data points for comparison for Portuguese experiments, and 120 data points for Marathi.

After collecting data of the subjective task, we simply counted the total percentage of votes received by each model in an experiment. The model that receives the majority vote can be thought of as the winning model.

## 5. EXPERIMENTAL RESULTS

Our baseline phrasing model for all languages is the punctuation based model. However, text used in these experiments does not have punctuation in it, and hence our baseline model does no phrasing at all. We call this model the NONE model.

### 5.1. English Phrasing

Between the three languages we worked with, English is the one with the most resources available to us. We have four phrasing models for English: (i) The NONE model, (ii) Festival's standard phrasing model[2], (iii) A Grammar based phrasing model based on POS tags, and (iv) A Grammar based phrasing model based on IPOS tags. Table 1 shows the results of objective evaluation of these four models. The results presented here are the average values after performing 10-fold cross validation.

**Table 1**. Objective Results for Phrasing in English

| System | F1 | L2 | EMD |
|---|---|---|---|
| NONE | 0.0000 | 0.2566 | 10.6233 |
| Festival | 0.3417 | 0.2802 | 3.0733 |
| POS Phrasing | 0.3481 | 0.1661 | 1.1449 |
| IPOS Phrasing | 0.2751 | 0.1972 | 1.7744 |

Based on the results in Table 1 and performing significance analysis, we can draw the following conclusions for p-value $p < 0.01$:

- Grammar based POS phrasing model is slightly better than the default model in Festival. The improvement in F-1 measure is not significant, but the improvement in L2 and EMD measures is significant.
- Grammar based IPOS phrasing model is slightly weaker than the Grammar based POS model across all metrics, but the differences are not statistically significant.
- Both the IPOS and POS models are significantly better than the NONE model.

We wanted to see if subjective listening tests support the objective comparisons here. We did two listening tests. First, we compared the NONE model to the IPOS model. Table 2 shows the results for this. We found that the IPOS model gets more votes than the NONE model. The result is statistically significant.

**Table 2**. Subjective Results (1 of 2) for Phrasing in English

| | % Votes |
|---|---|
| Model "NONE" Better | 36.6% |
| Model "IPOS" Better | 56.1% |
| No Difference | 7.3% |

In the second test, we compared the IPOS model to the POS model. Table 3 shows these results. We found that while the POS model gets more votes overall compared to the IPOS model, the difference is not statistically significant.

**Table 3**. Subjective Results (2 of 2) for Phrasing in English

| | % Votes |
|---|---|
| Model "IPOS" Better | 42.5% |
| Model "POS" Better | 50.0% |
| No Difference | 7.5% |

We thus see that using the Grammar based approach with POS tags helps us do better at phrasing than the standard model in Festival. We also see that replacing the POS tagger with IPOS tags also gives us a very reasonable phrasing model.

### 5.2. Portuguese Phrasing

For Portuguese, we only have three phrasing models: (i) The NONE model, (ii) The Grammar based POS model, and (iii) The Grammar based IPOS model. Note that the POS model here is slightly different than the one available for English, because we only had a lexical part of speech available for Portuguese. Table 4 summarizes the objective results for Portuguese phrasing. The results presented here are average values after performing 10 fold cross validation.

**Table 4**. Objective Results for Phrasing in Portuguese

| System | F1 | L2 | EMD |
|---|---|---|---|
| NONE | 0.0000 | 0.4113 | 28.2284 |
| IPOS Phrasing | 0.2870 | 0.2427 | 2.9735 |
| POS Phrasing | 0.2520 | 0.2639 | 3.2327 |

After performing significance analysis over these objective results, we found like just like for English, we could make the following conclusions:

- Both the IPOS and POS models are significantly better than the NONE model.
- The IPOS model is not significantly different compared to the POS model.

We tried to verify with listening tests, whether these hypotheses hold true for subjective opinion also. We did two listening tests, similar to those in English.

In the first listening test, we compared the NONE model to the IPOS model. Table 5 lists the results. In the second test, we compared the IPOS model to the POS model. Table 6 shows that result. Numerically, we see that the IPOS model is better than the NONE model, and that the POS model is better than the IPOS model. Significance analysis showed that the three systems may not be significantly different on the listening tasks.

**Table 5**. Subjective Results (1 of 2) for Phrasing in Portuguese

| | % Votes |
|---|---|
| Model "NONE" Better | 27.7% |
| Model "IPOS" Better | 46.8% |
| No Difference | 25.5% |

**Table 6**. Subjective Results (2 of 2) for Phrasing in Portuguese

| | % Votes |
|---|---|
| Model "IPOS" Better | 36.7% |
| Model "POS" Better | 50.0% |
| No Difference | 13.3% |

### 5.3. Marathi TTS

We only have two phrasing models for Marathi: (i) The NONE model, and (ii) A Grammar based model trained with IPOS tags. Table 7 summarizes the average objective results after 10 fold cross validation, comparing these two models, and subjective results are presented in table 8. We see that both objectively and subjectively, the IPOS model is significantly ($p < 0.01$) better than not having phrasing at all.

**Table 7**. Objective Results for Phrasing in Marathi

| System | F1 | L2 | EMD |
|---|---|---|---|
| NONE | 0.0000 | 0.1850 | 2.1491 |
| IPOS Phrasing | 0.2560 | 0.1828 | 0.8352 |

## 6. CONCLUSIONS AND FUTURE WORK

In this paper, we described how to build grammar based phrasing models for new languages. We use a data driven approach, by learning to predict acoustically derived phrases with the help of word categories such as part of speech, either provided by a tagger, or induced automatically from data.

We induced parts of speech over small corpora (50k sentences) for English, Portuguese and Marathi and built phrasing models using these tags. We also trained models for English and Portuguese using linguistically motivated parts of speech. Our results show that automatically derived IPOS tags yield models close to those obtained with POS tags.

When synthesizing low resource languages, we can build significantly better voices by using these grammar based models, rather than relying on the fragile punctuation based baselines. Investing time and money in building linguistic tools such as POS taggers for these languages may yield slightly better models, but the phrasing may not be remarkably different than the one obtained using automatic tags.

We had also looked at building phrasing models for Pashto, a language spoken in Afghanistan. We noticed that the speech data available to us for building a voice deliberately had short utterances, and there were no breaks in the recordings to train phrasing models from. We would like to investigate using the Pashto ASR training data that has longer utterances to train phrasing models.

**Table 8**. Subjective Results for Phrase Prediction in Marathi

| | % Votes |
|---|---|
| Model "NONE" Better | 22.5% |
| Model "IPOS" Better | 57.5% |
| No Difference | 20.0% |

## 7. ACKNOWLEDGMENT

This work was supported by the Fundação de Ciência e Tecnologia through the CMU/Portugal Program, a joint program between the Portuguese Government and Carnegie Mellon University.

## 8. REFERENCES

[1] Alan W Black and Paul Taylor, "The festival speech synthesis system: system documentation," Tech. Rep., Human Communication Research Centre, University of Edinburgh, January 1997.

[2] Paul Taylor and Alan W Black, "Assigning phrase breaks from part-of-speech sequences," *Computer Speech and Language*, vol. 12, pp. 99–117, 1998.

[3] P. Roach, G. Knowles, T. Varadi, and S. Arnfield, "MARSEC: A machine-readable spoken english corpus," *Journal of the International Phonetic Association*, vol. 23, no. 1, pp. 47–53, 1993.

[4] Mari Ostendorf, Patti J. Price, and Stefanie Shattuck-Hufnagel, "The boston university radio news corpus," Tech. Rep., Boston University, March 1995.

[5] Philipp Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Machine Translation Summit*, Phuket, Thailand, September 2005, pp. 79–86.

[6] Alok Parlikar and Alan W Black, "A grammar based approach to style specific phrase prediction," in *Interspeech*, Florence, Italy, August 2011, pp. 2149–2152.

[7] Kishore Prahallad, Alan W Black, and Ravishankhar Mosur, "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," in *ICASSP*, Toulouse, France, May 2006, vol. 1, pp. 853–856.

[8] Hermann Ney, Ute Essen, and Reinhard Kneser, "On structuring probabilistic dependences in stochastic language modelling," *Computational Linguistics*, vol. 8, no. 1, pp. 1–38, 1994.

[9] Alexander Clark, "Combining distributional and morphological information for part of speech induction," in *European Chapter of Association for Computational Linguistics*, Budapest, Hungary, August 2003, pp. 59–66.

[10] C. J. van Rijsbergen, *Information Retrieval*, Butterworth, 1979.

[11] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas, "A metric for distributions with applications to image databases," *International Conference on Computer Vision*, vol. 0, pp. 59, 1998.

[12] Scott Cohen, "Finding color and shape patterns in images," Tech. Rep., Stanford University, 1999.